DOCUMENT RESUME

ED 300 398 TM 012 373

AUTHOR Hambleton, Ronald K.; Rogers, H. Jane
TITLE Detecting Biased Test Items: Comparison of the IRT Area and Mantel-Haenszel Methods.
PUB DATE Apr 88
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; High Schools; High School Students; *Item Analysis; *Latent Trait Theory; *Statistical Bias; *Test Bias; Test Construction; Test Items; White Students
IDENTIFIERS *Item Characteristic Curve Area; *Mantel Haenszel Procedure; Native Americans; New Mexico High School Proficiency Examination

ABSTRACT

The agreement between item response theory-based and Mantel Haenszel (MH) methods in identifying biased items on tests was studied. Data came from item responses of four spaced samples of 1,000 examinees each--two samples of 1,000 Anglo-American and two samples of 1,000 Native American students taking the New Mexico High School Proficiency Examination in 1982. In addition, a matched group analysis was conducted using a third sample of 650 Native Americans and 650 Anglo Americans. The item characteristic curve area and the MH methods were used. The consistency of classification of items into biased and not-biased was in the 75 to 80% range for both methods. When the unreliability of item bias statistics was taken into account, both methods gave similar results. Discrepancies between methods were due to bias from intersections of item characteristic curves and the choice of interval over which item bias was defined. The Mantel-Haenszel method, with a minor modification or two, provides an acceptable approximation to the item response theory based methods. Five data tables and eight graphs show study results. (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED300398

# Detecting Biased Test Items: Comparison of the IRT Area and Mantel-Haenszel Methods

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts at Amherst

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald K. Hambleton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 012 373

BEST COPY AVAILABLE

# Detecting Biased Test Items: Comparison of the IRT Area and Mantel-Haenszel Methods[1,2]

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts at Amherst

## Abstract

IRT-based methods for identifying biased test items have considerable appeal, however, difficult problems sometimes arise in applying them. The Mantel-Haenszel method (MH) shares some of the desirable features of IRT-based item bias methods but not most of the difficulties. The main purpose of this study was to determine the degree of agreement between the IRT-based and MH methods in identifying biased items and, when the two methods led to different results to identify possible reasons for the discrepancies.

Data for the study came from the item responses of Anglo-American and Native-American students who were administered the 1982 New Mexico High School Proficiency Exam. Two samples of 1000 students from each group were used in the item bias analyses. Item bias methods studied were the ICC Area method (using 3-parameter ICCs) and the Mantel-Haenszel method.

The main findings were that (1) the consistency of classifications of items into biased and not-biased categories across replications was in the 75 to 80% range for both methods, and (2) when the unreliability of item bias statistics was taken into account, the two methods led to very similar results. Discrepancies between methods were due to bias resulting from intersecting ICCs (the Mantel-Haenszel method could not identify these items) and the choice of interval over which item bias was defined (the IRT method results depended on the choice of interval). The implications of the results for practitioners seem clear: The Mantel-Haenszel method, with a minor modification or two, provides an acceptable approximation to the IRT-based methods.

TM012 373

# Detecting Biased Test Items: Comparison of the IRT Area and Mantel-Haenszel Methods[1,2]

Ronald K. Hambleton and H. Jane Rogers
University of Massachusetts at Amherst

Recent attention to the detection of biased test items has resulted in the generation of a plethora of item bias methods (see, for example, Berk, 1982). Perhaps of greatest interest at present are IRT-based methods and the Mantel-Haenszel method. IRT-based methods have become popular and indeed, are considered "theoretically preferred" by some researchers (e.g., Shepard, Camilli, & Averill, 1981; Shepard, Camilli, & Williams, 1984) because of their close connection to the most widely accepted definition of item bias. This definition states that an item is biased if examinees of the same ability but from different sub-groups do not have the same probability of a correct response to the item. Thus the study of item bias within an IRT framework is a matter of comparing the item characteristic curves (ICCs) for the two sub-groups of interest (Hambleton & Swaminathan, 1985). Choice of mathematical form of the ICCs and approach to representing the differences between ICCs give rise to many of the IRT-based methods.

The Mantel-Haenszel method, proposed by Holland and Thayer (1986, 1988), also compares the probabilities of a correct response

---

[1]A paper presented at the annual meeting of AERA, New Orleans, 1988.

[2]Laboratory of Psychometric and Evaluative Research Report No. 175. Amherst, MA: School of Education, University of Massachusetts, 1988.

in the two groups of interest for examinees of the same ability, although its calculation is very different from the IRT-based methods (Holland & Thayer, 1988).

While IRT-based methods may have theoretical appeal, they have several drawbacks in practice, particularly when the three-parameter IRT model is used. High costs associated with running an IRT computer program such as LOGIST, large sample requirements, and sometimes poor parameter estimates, make implementation of IRT item bias methods problematic, if not impossible in some situations. Moreover, careful attention must be given to the scaling of item parameters, the choice of ability interval over which bias is measured, and (sometimes) the determination of a "cutoff" value for interpreting the results. The Mantel-Haenszel method, on the other hand, shares some of the desirable features of the IRT methods but not most of the difficulties. Computer programs for calculating the statistic are easily written; the cost of the analysis is low (probably under ten dollars); sample sizes need not be as large as for IRT-based methods; and a significance test is available to aid in interpreting the bias statistics. This simplicity is achieved at the expense of some generality, however; the calculation of a Mantel-Haenszel statistic may be considered analogous to comparing two item characteristic curves based on a one-parameter logistic model. Thus the Mantel-Haenszel statistic is not designed to detect non-uniform item bias.

Two different arguments can be used to support an interest in the Mantel-Haenszel method. For IRT advocates, the Mantel-Haenszel method may seem an acceptable approximation. For others, the Mantel-Haenszel method may seem preferable because of the logic underlying the method, its conceptual simplicity, and the availability of suitable significance tests. In view of the wide acceptance of IRT-based methods and the current interest in the Mantel-Haenszel method, a comparison of the results when applied to the same test data seemed timely. A previous study by Hambleton, Rogers, and Arrasmith (1988) provided some initial findings of the high agreement between the Mantel-Haenszel and one of the IRT-based methods (ICC Area method) when methodological problems associated with the methods were taken into account. The desirability of repeating their study with other datasets was noted by the authors.

## Purposes

The main purpose of this research was to carry out a detailed analysis of the item bias results obtained from an IRT-based item bias method and the Mantel-Haenszel method. Specifically, interest was centered on the degree of agreement between the methods in identifying biased items, and on possible reasons for disagreements when they were found. The research was primarily intended to determine the consequences of substituting an IRT-based item bias method for the easier to use and more convenient Mantel-Haenszel method.

A second purpose of the study was to examine the behavior of the item bias statistics when the ability distributions of the two groups of interest are considerably different. Widely differing ability distributions can be expected to affect the quality of IRT parameter estimation in one or both groups, and hence will influence the values of the IRT-based item bias statistics. The effect of discrepant score distributions on the Mantel-Haenszel statistic is more difficult to predict, hence it was of interest to study the situation.

## Method

### Description of the Test Data and Examinee Samples

The samples used in the study were drawn from a dataset containing the responses of approximately 23,000 students to the 1982 New Mexico High School Proficiency Exam (NMHSPE). The NMHSPE is a 150-item test which assesses "life skills" in five major areas: Knowledge of Community Resources, Consumer Economics, Government and Law, Mental and Physical Health, and Occupational Knowledge. Of the total group of students, approximately 8,000 were Anglo-American and 2,600 were Native American. This dataset was chosen for the study because of the widely discrepant score distributions of the Anglo- and Native Americans, and because of the large number of items flagged as potentially biased in an earlier item bias investigation (Hambleton, Martois, & Williams, 1983).

Description of the Item Bias Statistics

*ICC Area Method.* The ICC Area method entails the calculation of the area between the item characteristic curves obtained for each group separately (Rudner, Getson, & Knight, 1980). The area is calculated over a specified ability interval, which in this study was from the lower group mean minus three standard deviations to the upper group mean plus three standard deviations. Because there is no known sampling distribution for the area statistic under the null hypothesis of no group differences, items are typically ranked according to the values of the statistic and those with the highest values flagged as potentially biased. In this study, a "cutoff" value was obtained by carrying out an analysis on two randomly equivalent groups (the two Native American samples). Since there is no bias present, the largest area statistic obtained serves as an indicator of the greatest value of the statistic likely to occur by chance (for a further discussion, see Rogers & Hambleton, in press). This approach is not ideal; however, it does provide an approximate answer to the cut-off score determination problem.

*The Mantel-Haenszel Statistic.* The Mantel-Haenszel method works directly with the item responses for the two groups (referred to in the psychometric literature as the reference group and the focal group). As described earlier, examinees are first sorted into score groups according to total test score, resulting in up to (n + 1) score groups. Within the jth score group, a 2x2 table of frequencies is set up:

| | | Item Score 1 | Item Score 0 | |
|---|---|---|---|---|
| Group | Reference | $A_j$ | $B_j$ | $n_{Rj}$ |
| | Focal | $C_j$ | $D_j$ | $n_{Fj}$ |
| | | $m_{1j}$ | $m_{0j}$ | $T_j$ |

$A_j$, $B_j$, $C_j$, and $D_j$ correspond to the numbers of examinees in the four cells of the 2x2 Table: $n_{Rj}$, $n_{Fj}$, $m_{1j}$, and $m_{0j}$ are the marginals. $T_j$ is the number of examinees in the jth score group who attempted the item under investigation. The Mantel-Haenszel Chi-Square Test Statistic has the form:

$$\frac{(|T_jA_j - T_jE(A_j)| - 2)^2}{T_jVar(A_j)}$$

where $E(A_j) = n_{Rj}\, m_{1j}/T_j$

and

$$Var(A_j) = \frac{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}{T_j^2(T_j-1)}$$

(From Holland & Thayer, 1988).

The Mantel-Haenszel statistic tends to be large, an indicator of item bias, when item performance in the reference and focal groups over the (n + 1) score groups is consistently different. For example, if the reference group outperforms the focal group by 10% on the average across the (n + 1) score groups, the MH chi-square test statistic will be large, and the correct interpretation is that the item is biased against the focal group.

Computer Programs

*LOGIST.* To estimate item characteristic curves, the LOGIST program (Wood & Lord, 1976) was used. LOGIST estimates parameters using the method of maximum likelihood. A modified Newton's method is used to solve the likelihood equations. Estimation is conducted in stages, in which first the item parameters are held fixed while ability parameters are estimated, then the ability estimates obtained are held fixed while item parameters are estimated.

In this study, three-parameter item characteristic curves were fitted to all items for each sample.

*IRTBIAS.* This FORTRAN V program, written by the second author, calculates the area between item characteristic curves as described earlier. Output from the LOGIST program for the two groups of interest may be input directly into IRTBIAS. The two sets of b-parameter estimates are first placed on a common metric by scaling both to a mean of zero and standard deviation of one. The other item parameter estimates and ability estimates are then transformed accordingly (Hambleton & Swaminathan, 1985). Item characteristic curves are calculated and the area between them computed over an ability interval specified by the user.

Output from the program includes the value of the total area between the ICCs for each item as well as the values of the "positive" and "negative" areas, i.e., the area for which the ICC for

the reference group is higher than that of the focal group, and vice versa.

*MH* STATISTIC. This program, also written by the second author in FORTRAN V, calculates the Mantel-Haenszel item bias statistic. By default, (n+1) score groups are constructed, where n is the number of items.

Mantel-Haenszel statistics are computed in two steps, as recommended by Holland (1985). First, score groups are constructed using total scores based on all items. Mantel-Haenszel statistics are then calculated for all items. Those items with Mantel-Haenszel values exceeding the tabulated chi-square value at the .01 level of significance are identified. Next, total scores are recalculated excluding these items. With this "purified" criterion for ability, score groups are reformed and the Mantel-Haenszel statistics are computed once more.

Output from the program includes frequency distributions for the two sub-groups, and results of the analysis for the first and second steps. For each item, p-values for each sub-group, the common odds-ratio, the Mantel-Haenszel chi-square statistic, and the corresponding z-value are printed.

Procedure

For the purposes of the study, four spaced samples of 1000 examinees each were drawn: two samples were Anglo-American and two were Native American. To facilitate the analysis and reduce computer

time, only 75 of the 150 items were used. Items were chosen such that very easy items ($p > .90$) and items with very low discrimination ($r < .1$) were excluded; such items cause difficulties in IRT parameter estimation and often lead to unusually unstable item bias statistics. Three-parameter IRT models were fitted to each of the four samples separately.

With two Anglo-American and two Native American samples, two independent bias analyses could be carried out. The second comparison was conducted to enable examination of the consistency with which each bias statistic flagged items across samples. The Mantel-Haenszel and ICC Area method statistics were calculated for each item in each of the two comparisons.

To study the effect of the discrepancy in the score distributions of the two groups on the Area method statistics, a variation of the statistic was also computed. The ability interval over which the area is calculated was modified to cover the ability scale from two standard deviations below the Native American group mean to two standard deviations above the same mean. By restricting the interval in this way, attention was focused on that part of the ability scale where most of the Native American examinees were located, and hence where differences between the Anglo- and Native American ICCs were of greatest practical significance.

To study the effect of the score distribution differences on the Mantel-Haenszel statistic, a matched group analysis was carried out. For this analysis, a third sample of Native Americans was selected

such that the distribution of scores more closely matched that of the Anglo-American sample. A sample of 650 Native Americans was obtained and compared with a sample of 650 Anglo-Americans. Both the Mantel-Haenszel and Area method statistics were calculated.

## Results

As can be seen from Figures 1 and 2, the score distributions for Anglo- and Native Americans were considerably different. The mean and standard deviation for the first Anglo-American sample were 54.23 and 10.65 respectively, and for the first Native American sample, the values were 36.55 and 11.19, respectively. The mean and standard deviation for the second Anglo-American sample were 54.21 and 10.61 respectively, and for the second Native American sample, the values were 37.01 and 11.86, respectively.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Insert Figures 1 and 2 and Table 1 about here.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - -

After obtaining three-parameter model estimates for examinees and items for the four samples (two Anglo-American, two Native American), absolute-valued standardized residuals were calculated to determine the appropriateness of the fits between the model and the test data (Hambleton & Swaminathan, 1985). Table 1 provides a summary of the results. The results show clearly that there was a very close match between the three-parameter model and each set of test data (see, for example, Hambleton & Rogers, in press). The results were slightly better for the Anglo-American samples, but the fits were excellent for all four datasets.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Insert Tables 2, 3, and 4 about here.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Using the two independent Anglo- vs. Native American comparisons, the consistency with which the Area method and the Mantel-Haenszel method flagged items as potentially biased was examined. Table 2 lists the items which were flagged by the Area method in one or both comparisons, and indicates those items which were consistently identified. The cut-off score for this method was .468. Table 3 reports similar information for the Mantel-Haenszel method. The cut-off score for this method was 6.60. Where an item was flagged in one sample but the statistic was borderline in the other sample, the result was treated as consistent. The consistency results for the two methods are summarized in Table 4.

From Tables 2 and 3, it can be seen that both methods displayed considerable instability across samples. Using the Area method, 20 of the 75 items were flagged in one comparison but not the other. Using the Mantel-Haenszel method, 15 items were inconsistently flagged. Overall consistency for the Area method was 73% and for the Mantel-Haenszel, 80%. This moderate level of consistency is somewhat surprising, considering that all of the results were based on 1000 examinees in each group, and disturbing in view of the fact that in most situations, the practitioner would not have the luxury of a cross-validation sample.

When the Area method and the Mantel-Haenszel method were compared, the results were more encouraging. This comparison, however, was carried out with items which were <u>consistently identified</u> as biased across samples with the same method. Table 5 lists the items

and values of the bias statistics for the 16 items consistently flagged by one or both methods. Of the 14 items consistently identified by the Area method across the two comparisons and the nine items consistently identified by the Mantel-Haenszel, seven items were common. Thus, those items identified by the Mantel-Haenszel method were more or less a subset of those identified by the Area method.

- - - - - - - - - - - - - - - - - - - - - - - - - -
Insert Table 5 about here.
- - - - - - - - - - - - - - - - - - - - - - - - - -

Attention was then focused on the nine items consistently flagged by one bias method but not the other. From Table 5, it can be seen that two of the items consistently flagged by the Area method (items 57 and 102) were flagged in one comparison by the Mantel-Haenszel statistic. Hence, the discrepancy in results for these items may have occurred due to a Type II error with the Mantel-Haenszel statistic. Conversely, item 11, which was consistently flagged by the Mantel-Haenszel statistic, was flagged in one comparison by the Area method, suggesting a Type II error resulting from the use of this method.

- - - - - - - - - - - - - - - - - - - - - - - - - -
Insert Figures 3 to 8 about here.
- - - - - - - - - - - - - - - - - - - - - - - - - -

For the remaining six items, ICCs for the two groups were plotted and are displayed in Figures 3 through 8. Figures 3 through 7 show the ICCs for the five items which were detected by the Area method but not the Mantel-Haenszel method. For four of these items (items 28, 30, 92, and 129), the ICCs crossed markedly. It is thus

not surprising that the Mantel-Haenszel statistic did not detect these items, since it is not designed to detect non-uniform bias. The discrepancy for item 88 is less easy to explain. This item is potentially biased against the Anglo-American sample. Over the interval (-3.7 to 3.5) the ICCs are clearly different and hence the item was flagged by the Area method. The MH method did not detect the item as biased because few Anglo-Americans scored in the region of the scale where the largest differences were observed. The mean ability score for the Anglo-Americans was about .90 (.91 in the first sample and .89 in the second); the standard deviation of ability scores was about .86 (.87 in the first sample and .85 in the second). Clearly, only a small percent of Anglo-Americans (perhaps about 15% were in the region of the scale where differences were observed.

Figure 8 shows the ICCs for item 60, which was flagged by the Mantel-Haenszel but not the Area method. The curves are uniformly but not markedly different. Since the most pronounced differences were in the region on the ability scale where many Native American examinees scored, it was likely that the Mantel-Haenszel method would detect item 60 as biased. The average ability score in the two Native American samples was about -.60 and the standard deviation was about 1.04. In contrast, the Area method addressed the differences in the ICCs over a much wider interval on the ability scale (-3.7 to 3.5). Over the full scale, the differences were relatively modest and therefore the item was not identified by the Area method.

When the Area statistic was computed over the restricted interval (Native American mean score $\pm$ 2 standard deviations, which was, approximately, -2.7 to 1.5), some changes in the ranking of items were observed. However, all but two of the items (items 92 and 102) that were consistently identified in the original analysis were consistently flagged over the narrowed and lower interval. Item 92 was an item where the ICCs crossed in the modified range, and did not diverge widely within the interval. Other items for which the ICCs crossed and which had large area values in the original analysis also tended to be ranked lower in the modified analysis. This result suggested that the modified area statistic might be more closely related to the Mantel-Haenszel results than the original area statistic for the dataset used in this study. When rank-order correlations were calculated, this proved to be the case. In sample 1, the rank-order correlation between the original area statistic and the Mantel-Haenszel statistic was .32, and between the modified area statistic and the Mantel-Haenszel, .48.

When the matched sample analysis was carried out, the Mantel-Haenszel results changed very little. All items which were previously consistently identified by the Mantel-Haenszel method were flagged again, along with four others that had been flagged in the sample 1 analysis. One item (item 28) was flagged which had not

previously been flagged in either comparison. The Area method results showed greater change, as might be expected in view of the reduction in sample size. Only five of the 14 items previously identified were flagged in this analysis.

## Discussion

Several major points emerge from these results. First, both IRT-based item bias methods and the Mantel-Haenszel method are somewhat unreliable in identifying biased items. The Area method results were consistent across samples (of 1000) about 73% of the time; the Mantel-Haenszel results were consistent about 80% of the time. This finding reinforces our preference for considering items only "potentially" biased on the basis of the value of the bias statistic. Also, this result helps to explain the moderate agreement reported in the measurement literature among item bias methods concerning items flagged as potentially biased. The fact is that studies of convergence of item bias methods are influenced greatly by the unreliability of item bias statistics.

Second, there is substantial agreement between an IRT-based item bias method (the Area method) and the Mantel-Haenszel method in the detection of uniformly biased test items. Also, the IRT-based item bias method appears to detect non-uniformly biased items; the Mantel-Haenszel method does not.

When the interval over which the area statistic was calculated was changed, the rankings of items according to the value of the

statistic also changed. Restricting the ability interval to focus on the region of the scale where most of the focal group is distributed may lead to the identification of fewer non-uniformly biased items with the Area method, and hence, greater congruence with the Mantel-Haenszel results. Of course, in practice, the choice of interval over which item bias is measured is an important methodological consideration, and must be considered when interpreting item bias statistics.

The distribution of test scores appears to have little impact on the Mantel-Haenszel results. Matching the groups according to test score distribution before calculating the bias statistics did not substantially change the results. The Area method results were influenced to a much greater extent, although this may have been due in part to the reduction in sample size which was necessary to achieve matching. In any case, the Mantel-Haenszel method showed much greater stability in the face of reduced sample size than did the Area method.

The implications of the results of this study for practice seem clear. First, practitioners should be reminded about the unreliability of item bias statistics. This means that they should be encouraged to use large samples in their analyses whenever possible and interpret item bias statistics with a fair degree of caution. Second, the evidence suggests that the Mantel-Haenszel method can be safely substituted for IRT-based methods if safeguards are put in place to detect non-uniformly biased items. These items

are likely to go undetected by the Mantel-Haenszel method. One safeguard would be to routinely compare the direction of the difference in p-values for the two groups of interest across score groups. If the direction of the difference favored one group at test scores below a certain test score and favored the other group above the test score, non-uniform bias could be suspected. Test items showing this pattern of performance in the two groups, though not identified by the Mantel-Haenszel method, could also be studied for possible bias. Analyses like the one proposed can easily be incorporated into computer programs to carry out the Mantel-Haenszel method and provide some protection across non-uniform biased items going undetected. Other simple safeguards, such as graphing techniques, could also be incorporated into the method to detect non-uniformly biased test items.

## References

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore, MD: The Johns Hopkins University Press.

Hambleton, R. K., Martois, J. S., & Williams, C. (1983, April). Detection of biased test items with item response models. Paper presented at the annual meeting of AERA, Montreal.

Hambleton, R. K., Rogers, H. J., & Arrasmith, D. (1988). Identifying potentially biased test items: A comparison of the Mantel-Haenszel statistic and several item response theory methods. Laboratory of Psychometric and Evaluative Research, Report No. 154. Amherst, MA: School of Education, University of Massachusetts.

Hambleton, R. K., & Rogers, H. J. (in press). Promising directions for assessing item response model fit to test data. Applied Psychological Measurement.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Holland, P. W. (1985). On the study of differential item difficulty. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1986). Differential item functioning and the Mantel-Haenszel procedure. Technical Report No. 86-31. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun, Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rogers, H. J., & Hambleton, R. K. (in press). Evaluating computer-simulated baeline statistics for interpreting item bias statistics. Educational and Psychological Measurement.

Rudner, L. M., Getson, P. P., & Knight, D. L. (1980). A Monte Carlo comparison of seven biased item detection techniques. Journal of Educational Measurement, 17, 1-10.

Shepard, L. A., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. Journal of Educational Statistics, 6, 317-375.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-138.

Wood, R. L., & Lord, F. M. (1976). A user's guide to LOGIST. Research Memorandum. Princeton, NJ: Educational Testing Service.

Figure 1. Test Score Distributions for the first Anglo-American (AA) and Native American (NA) sample.

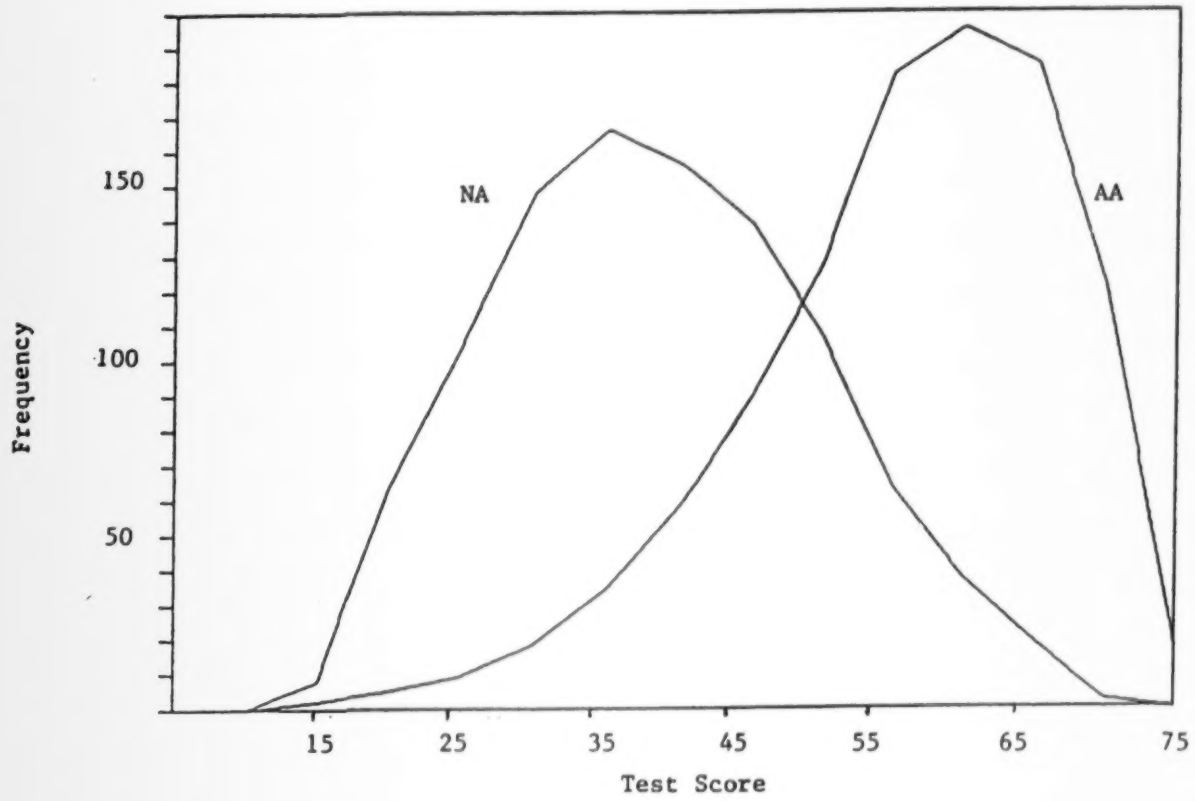


Figure 2. Test Score Distributions for the second Anglo-American (AA) and Native American (NA) sample.

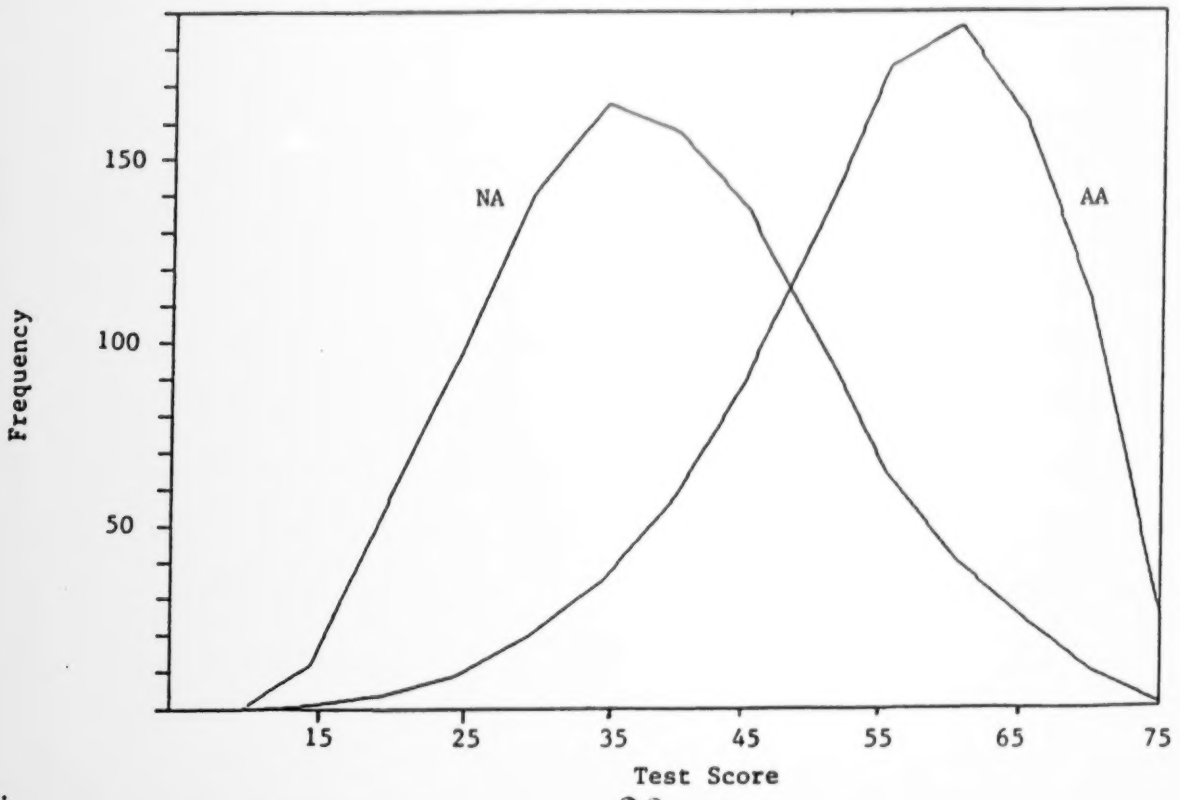

Figure 3. Anglo- and Native American ICCs for item 28.

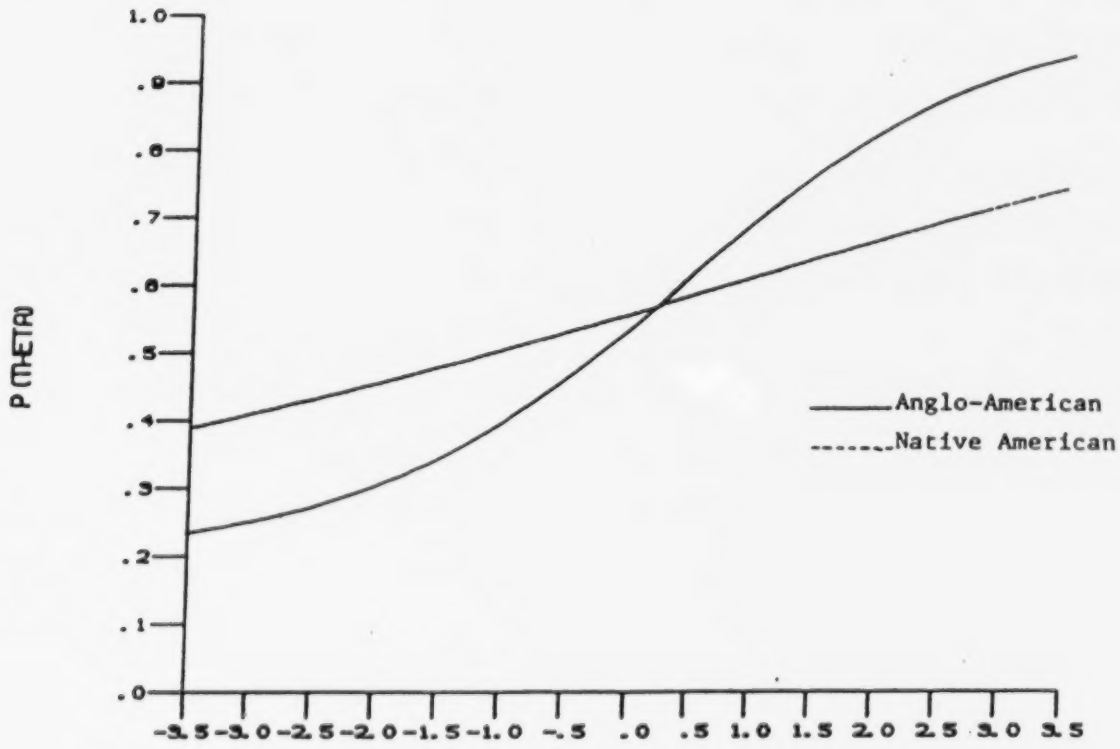


Figure 4. Anglo- and Native American ICCs for item 30.

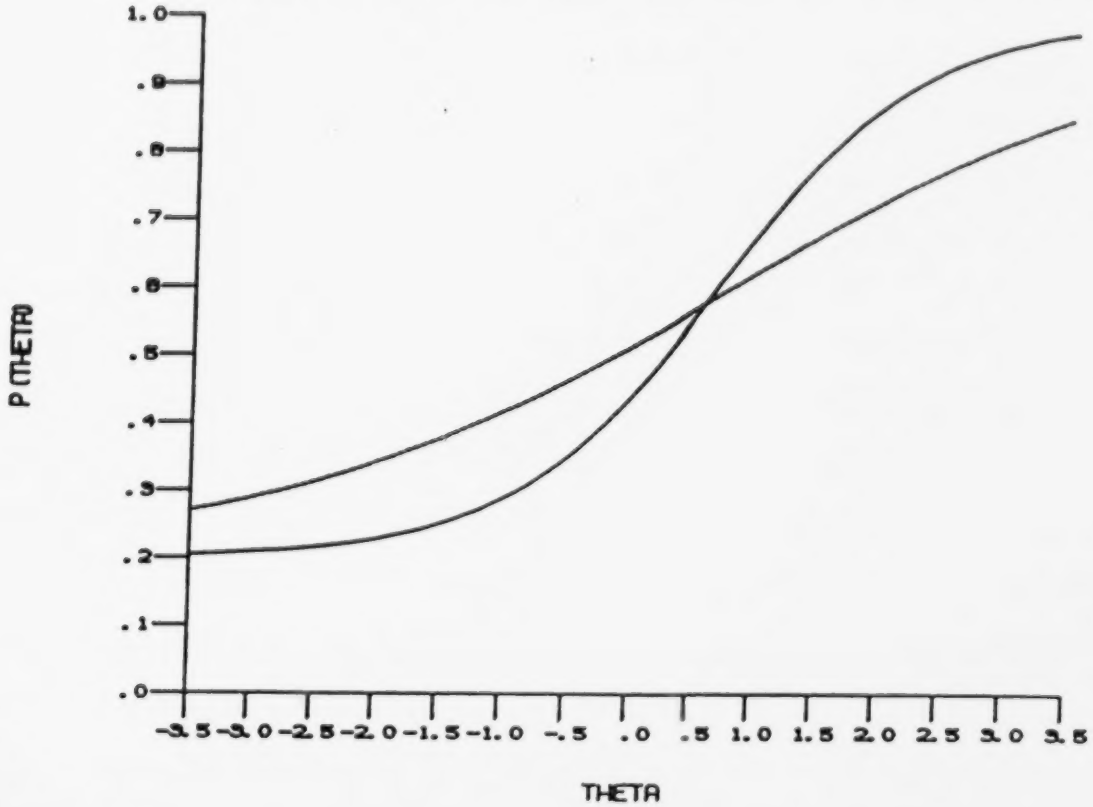

Figure 5. Anglo- and Native American ICCs for item 88.

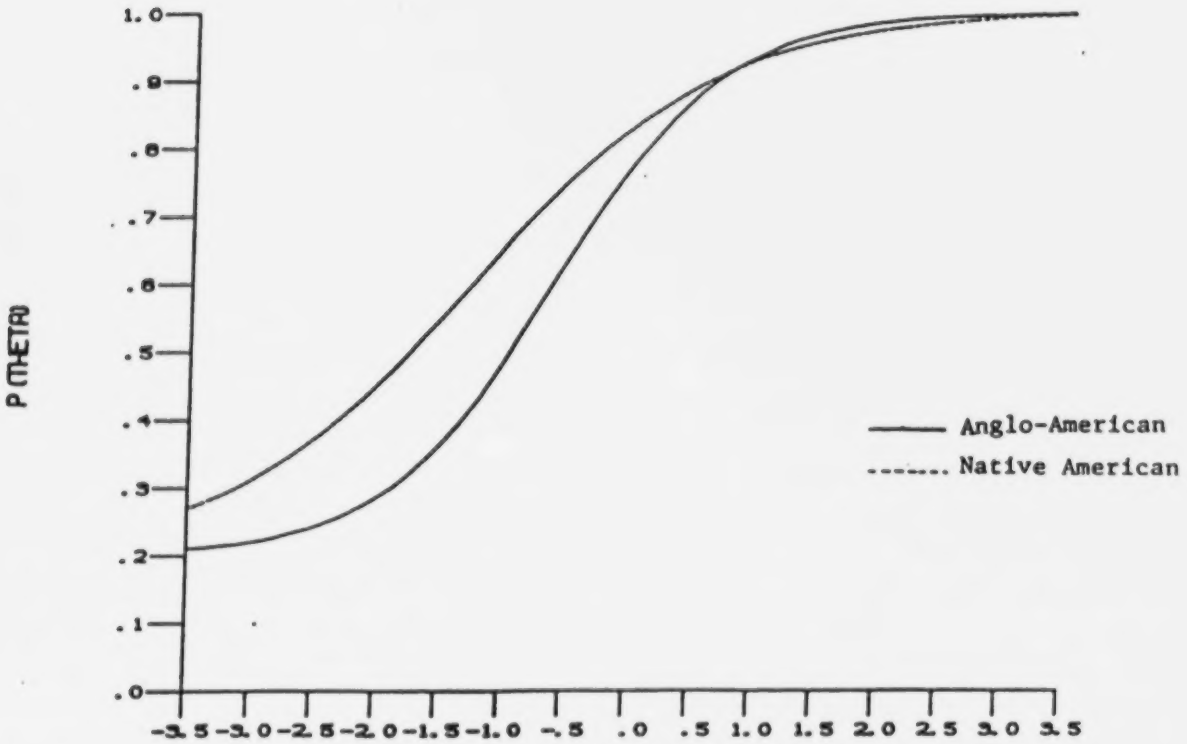


Figure 6. Anglo- and Native American ICCs for item 92.

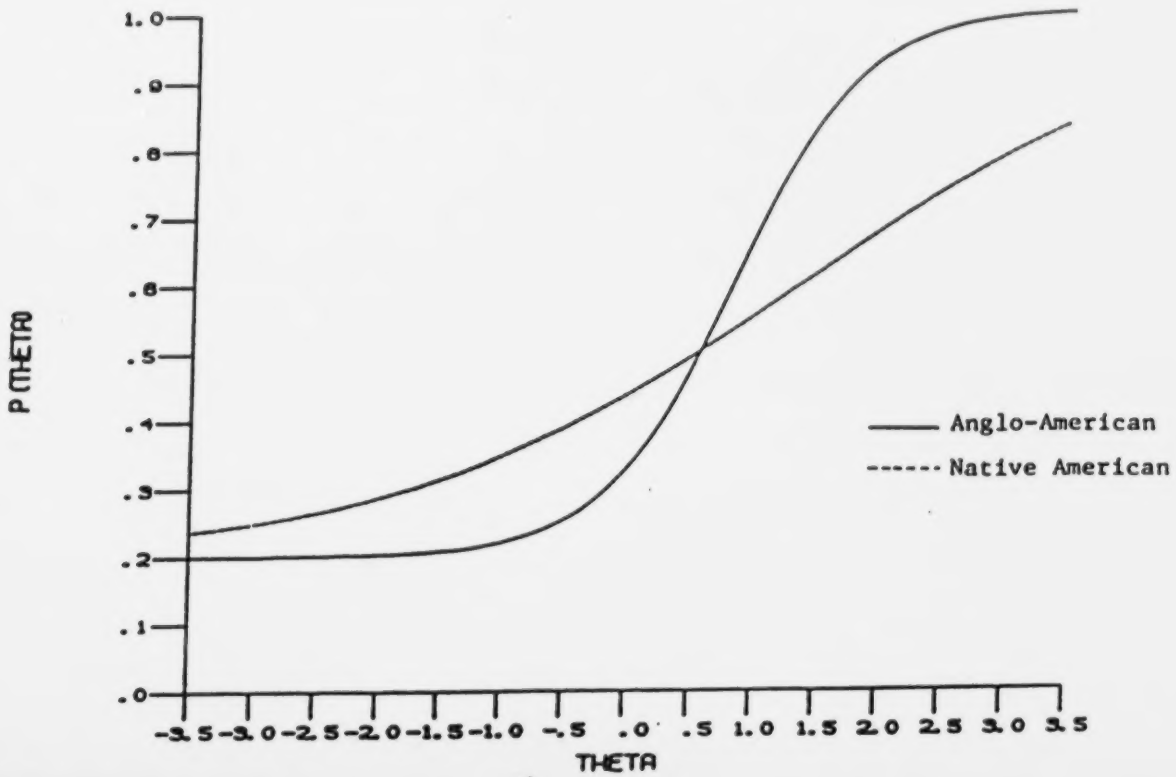

Figure 7. Anglo- and Native American ICCs for item 129.

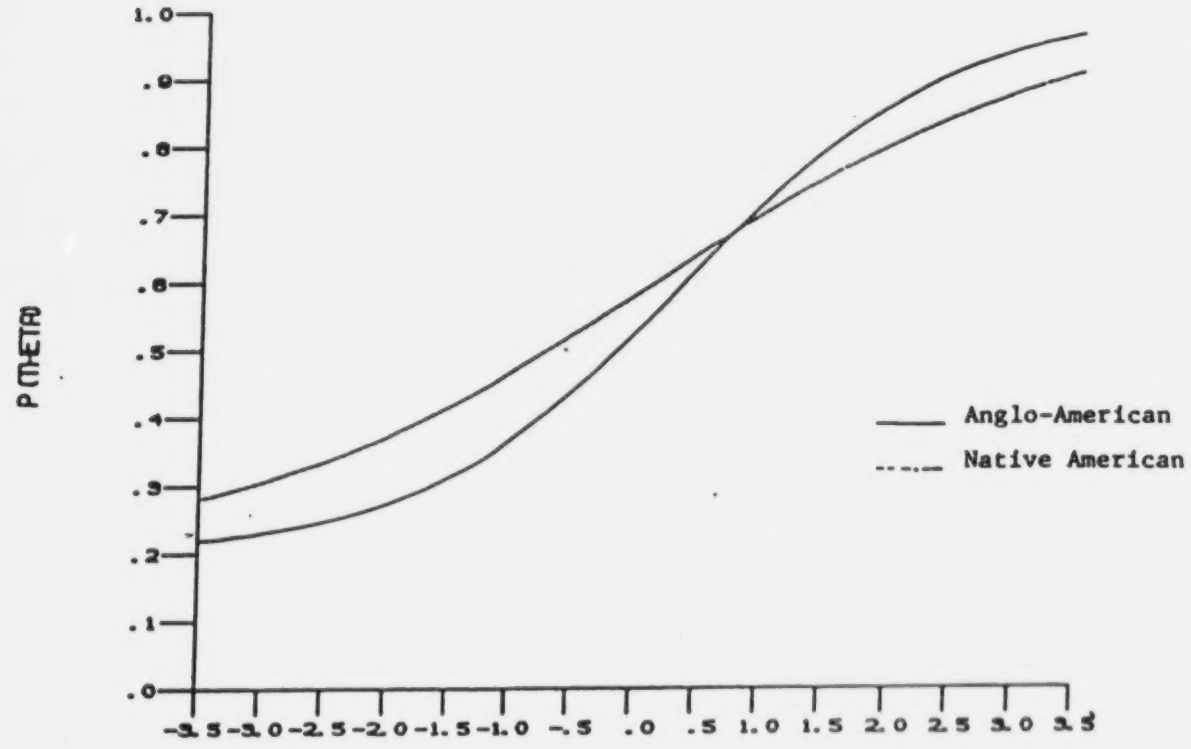


Figure 8. Anglo- and Native American ICCs for item 60.

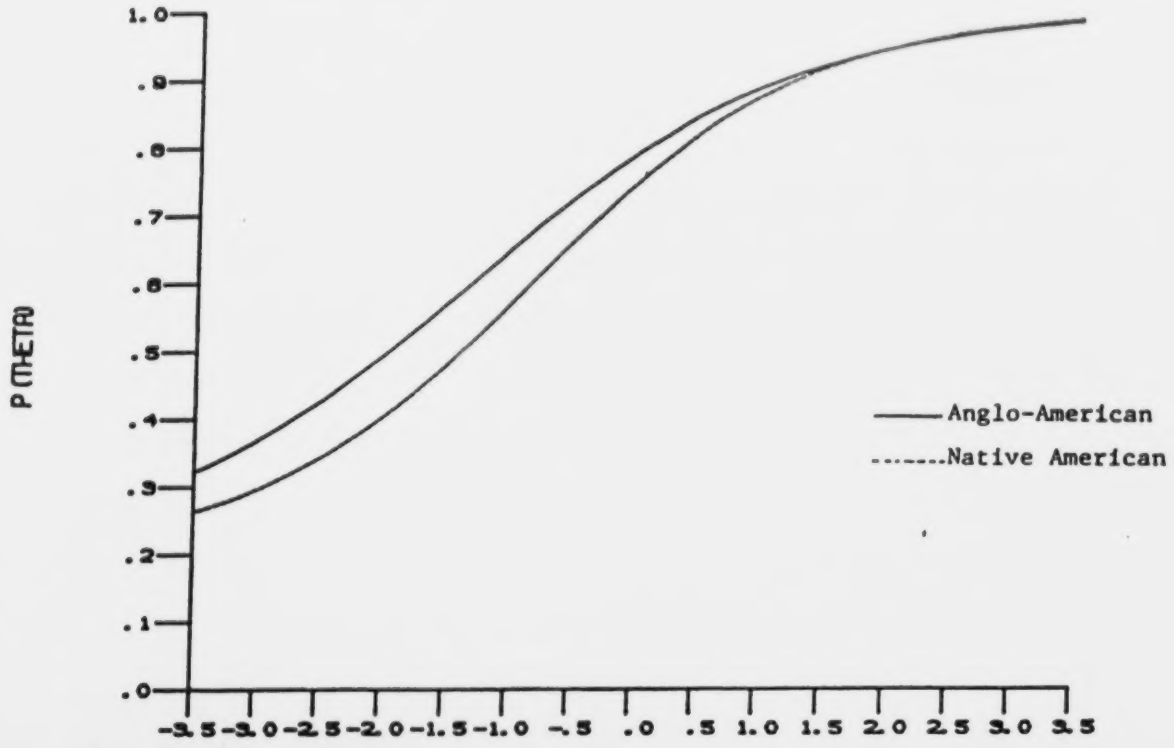

Table 1

Summary of the IRT Standardized Residuals

| Sample | Standardized Residuals \|0 to 1\| | \|1 to 2\| | \|2 to 3\| | Average |
|---|---|---|---|---|
| Anglo-American, Sample 1 | 72% | 25% | 3% | .74 |
| Anglo-American, Sample 1 | 73% | 22% | 5% | .76 |
| Native American, Sample 1 | 66% | 29% | 5% | .83 |
| Native American, Sample 2 | 70% | 24% | 6% | .80 |

Table 2

Item Statistics for Potentially Biased Test Items
Identified using the IRT Area Method[1]
(Anglo-Americans vs Native Americans, N = 1000)

| Test Item | Sample 1 AA b | AA a | NA b | NA a | Bias Statistic | Sample 2 AA b | AA a | NA b | NA a | Bias Statistic | Stability |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 11 | -0.43 | 0.77 | -0.00 | 0.98 | 0.354 | -0.85 | 0.54 | -0.03 | 0.81 | 0.645 | |
| 14 | -1.23 | 0.36 | -2.09 | 0.29 | 0.382 | -0.50 | 0.49 | -2.71 | 0.19 | 1.036 | |
| 15 | -1.32 | 0.42 | -0.99 | 0.81 | 0.566 | -0.81 | 0.68 | -0.98 | 0.86 | 0.215 | |
| 23 | 0.71 | 0.93 | 0.73 | 0.34 | 0.875 | 0.47 | 0.82 | 0.61 | 0.48 | 0.451 | |
| 28 | 0.41 | 0.47 | 0.81 | 0.16 | 0.903 | 0.45 | 0.42 | 0.64 | 0.20 | 0.657 | X |
| 30 | 0.69 | 0.72 | 0.81 | 0.32 | 0.736 | 0.58 | 0.79 | 0.47 | 0.37 | 0.701 | X |
| 31 | -0.33 | 1.04 | -0.63 | 0.56 | 0.520 | -0.65 | 0.76 | -0.56 | 0.75 | 0.069 | |
| 39 | -0.04 | 0.62 | -0.19 | 0.58 | 0.115 | 0.21 | 0.91 | -0.09 | 0.39 | 0.783 | |
| 42 | -0.34 | 0.54 | -0.78 | 0.60 | 0.201 | -0.51 | 0.51 | 0.02 | 0.83 | 0.536 | |
| 43 | -0.53 | 0.53 | -0.60 | 0.40 | 0.269 | -0.78 | 0.58 | -0.59 | 0.35 | 0.491 | |
| 50 | 1.81 | 0.55 | 1.49 | 0.94 | 0.488 | 1.50 | 0.85 | 1.56 | 0.90 | 0.225 | |
| 56 | -1.92 | 0.34 | -0.76 | 0.50 | 0.647 | 0.25 | 1.14 | 0.42 | 0.69 | 0.195 | |
| 57 | -1.44 | 0.50 | -0.74 | 0.83 | 0.584 | -1.35 | 0.53 | -0.69 | 0.86 | 0.562 | X |
| 67 | 2.11 | 0.84 | 1.76 | 0.76 | 0.277 | 2.32 | 0.87 | 1.61 | 0.81 | 0.509 | |
| 69 | -1.17 | 0.55 | -1.19 | 0.42 | 0.245 | -1.80 | 0.37 | -0.98 | 0.55 | 0.521 | |
| 75 | 0.87 | 0.91 | 1.66 | 0.53 | 0.608 | 0.89 | 0.90 | 1.25 | 0.87 | 0.281 | |
| 78 | 1.88 | 1.45 | 2.59 | 0.59 | 0.687 | 1.81 | 1.19 | 1.76 | 1.02 | 0.215 | |
| 82 | -0.50 | 0.58 | 0.17 | 0.61 | 0.509 | -0.66 | 0.46 | 0.03 | 0.76 | 0.626 | X |
| 88 | -0.62 | 0.90 | -1.27 | 0.62 | 0.516 | -0.77 | 0.87 | -1.34 | 0.56 | 0.493 | X |
| 92 | 0.86 | 1.00 | 1.37 | 0.44 | 0.686 | 0.82 | 1.18 | 1.38 | 0.37 | 0.916 | X |

[1]Although a three-parameter model was fitted to the data, the c-parameters for all items reported here were estimated to be .20.

X designates test items which were identified as consistently potentially biased.

Table 2 (cont.)

Item Statistics for Potentially Biased Test Items
Identified using the IRT Area Method[1]
(Anglo-Americans vs Native Americans, N = 1000)

| Test Item | Sample 1 | | | | | Sample 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AA | | NA | | Bias | AA | | NA | | Bias | |
| | b | a | b | a | Statistic | b | a | b | a | Statistic | Stability |
| 93 | 1.44 | 1.06 | 1.22 | 1.33 | 0.354 | 1.60 | 1.24 | 1.09 | 1.79 | 0.534 | |
| 101 | 0.34 | 0.40 | 1.56 | 0.45 | 0.838 | 0.56 | 0.43 | 0.87 | 0.83 | 0.602 | X |
| 102 | 0.27 | 0.69 | -0.48 | 0.48 | 0.584 | -0.09 | 0.47 | -0.06 | 0.28 | 0.488 | X |
| 107 | -1.56 | 0.36 | -0.59 | 0.45 | 0.567 | -1.49 | 0.38 | -0.56 | 0.43 | 0.581 | X |
| 110 | -1.46 | 0.43 | -0.74 | 0.50 | 0.465 | -2.06 | 0.36 | -0.86 | 0.41 | 0.694 | X |
| 115 | 0.25 | 0.45 | -0.29 | 0.28 | 0.534 | 0.46 | 0.39 | -0.06 | 0.30 | 0.380 | |
| 118 | -1.56 | 0.28 | -2.26 | 0.30 | 0.396 | -2.02 | 0.28 | -2.96 | 0.29 | 0.485 | |
| 122 | -1.38 | 0.34 | -0.10 | 0.70 | 0.945 | -1.19 | 0.34 | -0.08 | 0.59 | 0.789 | X |
| 123 | -1.07 | 0.42 | -0.73 | 0.70 | 0.489 | -1.18 | 0.38 | -0.73 | 0.50 | 0.335 | |
| 125 | 0.77 | 0.21 | 0.98 | 0.30 | 0.322 | 0.90 | 0.18 | 0.84 | 0.44 | 0.751 | |
| 127 | -1.19 | 0.67 | -1.15 | 0.44 | 0.381 | -1.05 | 0.67 | -1.13 | 0.38 | 0.523 | |
| 128 | -0.59 | 0.64 | -0.10 | 1.28 | 0.617 | -0.73 | 0.56 | 0.05 | 1.16 | 0.732 | X |
| 129 | 0.40 | 0.56 | 0.17 | 0.35 | 0.477 | 0.37 | 0.67 | 0.79 | 0.24 | 0.941 | X |
| 130 | 1.67 | 0.32 | 0.55 | 0.61 | 0.747 | 1.77 | 0.36 | 0.71 | 0.50 | 0.577 | X |

[1]Although a three-parameter model was fitted to the data, the c-parameters for all items reported here were estimated to be .20.

X designates test items which were identified as consistently potentially biased.

Table 3

Item Statistics for Potentially Biased Test Items
Identified using the Mantel-Haenszel Method[1]
(Anglo-Americans vs Native Americans, N = 1000)

| Test Item | Sample 1 | | | | | Sample 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AA | | NA | | Bias | AA | | NA | | Bias | |
| | b | a | b | a | Statistic | b | a | b | a | Statistic | Stability |
| 11 | -0.43 | 0.77 | -0.00 | 0.98 | 17.49 | -0.85 | 0.54 | -0.03 | 0.81 | 20.56 | X |
| 14 | -1.23 | 0.36 | -2.09 | 0.29 | 0.33 | -0.50 | 0.49 | -2.71 | 0.19 | 6.83 | |
| 27 | 0.31 | 0.64 | -0.18 | 0.65 | 7.67 | 0.27 | 0.63 | -0.09 | 0.72 | 3.99 | |
| 35 | -0.22 | 0.97 | 0.20 | 1.75 | 25.83 | 0.03 | 1.35 | 0.19 | 1.49 | 4.65 | |
| 41 | 0.63 | 0.94 | 0.91 | 0.99 | 8.17 | 0.67 | 0.98 | 0.95 | 0.96 | 4.60 | |
| 47 | -0.37 | 1.19 | -0.64 | 1.34 | 0.62 | -0.20 | 1.56 | -0.67 | 1.14 | 8.66 | |
| 48 | -0.05 | 0.81 | 0.20 | 0.78 | 12.07 | 0.02 | 0.94 | 0.15 | 1.07 | 2.01 | |
| 56 | -1.92 | 0.34 | -0.76 | 0.50 | 12.10 | -0.95 | 0.46 | -0.88 | 0.57 | 0.13 | |
| 57 | -1.44 | 0.50 | -0.74 | 0.83 | 7.34 | -1.35 | 0.53 | -0.69 | 0.86 | 4.94 | |
| 60 | -1.32 | 0.46 | -0.84 | 0.54 | 8.08 | -1.06 | 0.56 | -0.95 | 0.39 | 7.30 | X |
| 64 | 0.23 | 1.06 | 0.22 | 0.74 | 0.02 | -0.00 | 0.86 | 0.34 | 0.85 | 9.52 | |
| 67 | 2.11 | 0.84 | 1.76 | 0.76 | 3.78 | 2.32 | 0.90 | 1.61 | 0.81 | 12.83 | |
| 75 | 0.87 | 0.91 | 1.66 | 0.53 | 7.01 | 0.89 | 0.46 | 1.25 | 0.87 | 3.42 | |
| 82 | -0.50 | 0.58 | 0.17 | 0.61 | 20.17 | -0.67 | 0.46 | 0.03 | 0.76 | 8.56 | X |
| 101 | 0.34 | 0.40 | 1.56 | 0.45 | 30.87 | 0.56 | 0.43 | 0.87 | 0.83 | 8.32 | X |
| 102 | 0.27 | 0.69 | -0.48 | 0.48 | 9.67 | -0.09 | 0.47 | -0.06 | 0.28 | 0.36 | |
| 104 | 0.82 | 0.85 | 0.24 | 1.02 | 14.94 | 0.67 | 0.84 | 0.39 | 0.80 | 5.99 | |
| 107 | -1.56 | 0.36 | -0.59 | 0.45 | 11.43 | -1.49 | 0.38 | -0.56 | 0.43 | 11.24 | X |
| 110 | -1.46 | 0.43 | -0.74 | 0.50 | 13.03 | -2.06 | 0.36 | -0.86 | 0.41 | 17.37 | X |
| 118 | -1.56 | 0.28 | -2.26 | 0.30 | 3.56 | -2.02 | 0.28 | -2.96 | 0.29 | 10.39 | |

[1]Although a three-parameter model was fitted to the data, the c-parameters for all items reported here were estimated to be .20.

X designates test items which were identified as consistently potentially biased.

Table 3 (cont.)

Item Statistics for Potentially Biased Test Items
Identified using the Mantel-Haenszel Method
(Anglo-Americans vs Native Americans, N = 1000)

| Test Item | Sample 1 | | | | | Sample 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AA | | NA | | Bias | AA | | NA | | Bias | |
| | b | a | b | a | Statistic | b | a | b | a | Statistic | Stability |
| 122 | -1.38 | 0.34 | -0.10 | 0.70 | 21.11 | -1.19 | 0.34 | -0.08 | 0.59 | 14.00 | X |
| 127 | -1.19 | 0.67 | -1.15 | 0.44 | 8.98 | -1.05 | 0.67 | -1.13 | 0.38 | 5.82 | |
| 128 | -0.59 | 0.64 | -0.10 | 1.28 | 19.50 | -0.73 | 0.56 | 0.01 | 1.16 | 16.27 | X |
| 130 | 1.67 | 0.32 | 0.55 | 0.61 | 7.49 | 1.78 | 0.36 | 0.71 | 0.50 | 12.59 | X |

[1]Although a three-parameter model was fitted to the data, the c-parameters for all items reported here were estimated to be .20.

X designates test items which were identified as consistently potentially biased.

Table 4

Summary of Results Concerning Consistency of
Bias - Non-Bias Classifications of 75 Test Items
in Two Independent Anglo- vs Native American Comparisons

| Category | Method | |
|---|---|---|
| | IRT Area | Mantel-Haenszel |
| Biased, Sample 1; Biased, Sample 2 | 14 | 9 |
| Biased, Sample 1; Non-Biased, Sample 2 | 9 | 10 |
| Non-Biased, Sample 1; Biased, Sample 2 | 11 | 5 |
| Non-Biased, Sample 1; Non-Biased, Sample 2 | 41 | 51 |
| Number of Consistently Classified Items | 55 | 60 |
| Percent of Consistently Classified Items | 73% | 80% |

Table 5

Agreement Between Methods in the Identification of Potentially Biased Test Items[1]

| Test Item | IRT Area Method S-1 | IRT Area Method S-2 | Mantel-Haenszel Method S-1 | Mantel-Haenszel Method S-2 | Agreement |
|---|---|---|---|---|---|
| 11 | (0.354)[2] | 0.645 | 17.49 | 20.56 | |
| 28 | 0.903 | 0.657 | (0.38) | (0.00) | |
| 30 | 0.736 | 0.701 | (0.10) | (4.54) | |
| 57 | 0.584 | 0.562 | 7.34 | (4.94) | |
| 60 | (0.315) | (0.349) | 8.08 | 7.30 | |
| 82 | 0.509 | 0.626 | 20.17 | 8.56 | X |
| 88 | 0.516 | 0.493 | (2.90) | (0.46) | |
| 92 | 0.686 | 0.916 | (0.01) | (0.11) | |
| 101 | 0.838 | 0.602 | 30.87 | 8.32 | X |
| 102 | 0.584 | 0.488 | 9.67 | (0.36) | |
| 107 | 0.567 | 0.581 | 11.43 | 11.24 | X |
| 110 | 0.465 | 0.694 | 13.03 | 17.37 | X |
| 122 | 0.945 | 0.789 | 21.11 | 14.00 | X |
| 128 | 0.617 | 0.732 | 19.50 | 16.27 | X |
| 129 | 0.477 | 0.941 | (2.11) | (0.41) | |
| 130 | 0.747 | 0.577 | 7.49 | 12.59 | X |

[1]Test items listed in the Table were consistently identified as biased by one or both methods.

[2]Values reported in brackets were <u>not</u> significant.